GPU BASED FAST PHYLOGENETIC TREE CONSTRUCTION ALGORITHM WITH REDUCE DATASET

NAJIHAH IBRAHIM

UNIVERSITI SAINS MALAYSIA

2016

GPU BASED FAST PHYLOGENETIC TREE CONSTRUCTION ALGORITHM WITH REDUCE DATASET

by

NAJIHAH IBRAHIM

Thesis submitted in fulfillment of the requirements
for the degree of
Master of Science

September 2016

ACKNOWLEDGEMENT

All the praise and thanks be to Allah SWT, the Most Beneficent and the Most Merciful.

First of all, I would like to express my deepest gratitude to my supervisor, Associate Professor Dr. Nur'Aini Abdul Rashid for her constant encouragement and guidance which have kept me always in the right track. Her support, motivation and comments have given me the strength that enables me to go through this challenge and learning process. I'm also grateful to other lectures, Professor Rosni Abdullah and Dr. Mohd Adib Haji Omar for the inputs and constructive insight towards this research.

Thank you to all School of Computer Sciences' lectures and staff for doing all the great work in managing and administering such a great environment for the entire student. Special thanks to Universiti Sains Malaysia that has partially support this research under the Research University (RU) Grant for "A GPU Based High Throughput Multiple Sequence Alignment Algorithm for Protein Data" [1001/PKOMP/817065] and thanks to Malaysia Government for the scholarship provided under MyBrain15 program.

Thank you to Ibrahim, Nadiah, Fazilah, Adilah, Ezzeddin, Hadri, Ramizah, Alfin, Syahmi, Nadzrin, Idzwan, Raed, Ahmad, Atheer, Awsan, Nizam, Jamal, Muhannad, Aisyah, Mubarak, Aszifa, Aimi, Afiqah, Asikin, Aini, Wawa, Mishal, Marwah, all the colleagues, friends and everyone at USM and PDCC lab for the fruitful discussions, guidance, moral support, encouragement and prayers that enlighten my way. My special thanks to my big family for their moral support and encouragement.

Last but not least, I would like to express my heartfelt gratitude and special regards to my mother, aunt and my late grandmother for their never-ending bonds and support, heart-warming feeling, strong believe and huge patience that always reminds me to fight hard in completing my study and make them proud in the end. To my beloved late grandparents, both of you will always be loved, forever.

"Victory Starts with Dreams, Faith and Determination"

Najihah Binti Ibrahim

Dreams.Hopes.Forward.Forever – BTS (The Most Beautiful Moment in Life)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

ALU	Arithmetic Logic Unit
BAliBASE	Benchmark Alignment Database
BEAST	Bayesian Evolutionary Analysis Sampling Trees
BLOSUM	BLOcks SUbstitution Matrix
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DDBJ	DNA DataBank of Japan
DNA	Deoxyribonucleic acid
DP	Dynamic Programming
EMBL	Europian Molecular Biology Laboratory
FFT	Fast Fourier Transform
GA	Genetic Algorithm
GAML	Genetic Algorithm for Maximum Likelihood
GARLI	Genetic Algorithm for Rapid Likelihood Inference
GenBank	National Institutes of Health Geanetic Sequence Database
GPU	Graphic Processing Unit
HIV	Human Immunodeficiency Virus
HKY	Hasegawa, Kishino and Yano
HOMSTRAD	Homologous Structure Alignment Database
HPC	High Performance Computing
JC	Jukes Cantor
M^3	Metropolis-Coupled Markov Chain Monte Carlo
MCMC	Markov Chain Monte Carlo

ML	Maximum Likelihood
MP	Maximum Parsimony
MSA	Multiple Sequence Alignment
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
NJ	Neighbor Joining
NNI	Nearest Neighbor Interchanges
OTU	Operational Taxonomic Unit
PA	Pairwise Alignment
PAM	Percent Accepted Mutations
P-GARLI	Parallel Genetic Algorithm for Rapid Likelihood Inference
Prefab	Protein Reference Alignment Benchmark
Pthreads	Posix Threads
RAM	Random-Access Memory
RNA	Ribonucleic Acid
SGI	Silicon Graphics International
SM	Streaming Multiprocessor
SPR	Sub-tree Pruning Regrafting
T-Coffee	Tree-based Consistency Objective Function For alignment Evaluation
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
WSP	Weighted Sum of Pairs

ALGORITMA PEMBINAAN POKOK FILOGENETIK YANG PANTAS MENGGUNAKAN GPU DENGAN PENGURANGAN DATASET

ABSTRAK

Kemajuan yang pesat dalam data genom yang baru, penambahbaikan dalam kaedah bagi menganalisis data genom, inovasi teknologi baru dan penyepaduan beberapa kaedah utama telah menjadi suatu kepentingan utama di dalam penyelidikan ini. Analisis jujukan telah digunakan untuk menganalisa dan memanipulasikan data genom yang homolog dan pokok filogenetik adalah salah satu kaedah yang telah digunakan dalam proses analisis jujukan. Pembinaan pokok filogenetik memerlukan proses awal iaitu proses penjajaran jujukan. Penjajaran jujukan adalah penting kerana set DNA data yang asal bagi kesemua spesis kebiasaannya adalah tidak berkualiti dan mempunyai aksara yang tidak dapat dikenalpasti. Penyelidikan ini telah membuktikan bahawa set data yang digunakan sebagai input mesti dijujukan terlebih dahulu sebelum pembinaan pokok filogenetik. Pada masa kini, terdapat pelbagai jenis atur cara yang boleh digunakan bagi melaksanakan penjajaran jujukan. Oleh itu, pemilihan atur cara yang sesuai untuk menjujukan set data sebagai input adalah sukar. Eksperimen awal yang dilakukan telah menunjukkan MAFFT sebagai atur cara yang terbaik untuk melaksanakan proses penjajaran jujukan berbanding ClustalW, Kalign, MUSCLE dan T-Coffee. Kajian perbandingan telah dilaksanakan untuk mendapatkan beberapa kaedah terbaik untuk pembinaan pokok filogenetik. Perbandingan telah dilakukan dengan membandingkan beberapa atur cara pembinaan pokok filogenetik yang terkemuka: GARLI, MrBayes, Tree Puzzle dan FastTree. FastTree telah dikenalpasti sebagai suatu atur cara yang mempunyai beberapa kaedah terbaik untuk membina pokok filogenetik, seperti kaedah penyatuan-jiran dan kaedah berdasarkan profil untuk

susunan nod dan posisi kategori. Melalui eksperimen tersebut, kami telah mengenalpasti bahawa input yang terbaik bagi pemilihan jujukan terjajar boleh mempengaruhi proses dan keputusan dalam pembinaan pokok filogenetik. Oleh itu, suatu kaedah telah diperkenalkan untuk meningkatkan kualiti jujukan terjajar tersebut. Penambahbaikan tersebut dikenali sebagai kaedah "Half-Parsimonious" dan berupaya untuk mengurangkan saiz dan mengekalkan bahagian yang berinformasi didalam set data tersebut. Kaedah "Half-Parsimonious" ini mampu meningkatkan skor bagi kebolehjadian maksimum dan ukuran dahan dan mengurangkan masa pemprosesan. Set data "Half-Parsimonious" akan digunakan sebagai input untuk proses penyepaduan beberapa kaedah untuk membina sebuah pokok filogenetik. Eksperimen kami telah menunjukkan bahawa algorithma bagi kaedah penyatuan baru ini mampu meningkatkan skor bagi kebolehjadian maksimum dan ukuran dahan. Walau bagaimanapun, masa pemprosesan bagi kaedah penyatuan baru ini telah meningkat akibat daripada perlaksaaan pencarian keseluruhan didalam algoritma tersebut. Oleh itu, kaedah pecutan telah dilaksanakan dengan menggunakan pemprosesan banyak-teras, iaitu Unit Pemprosesan Grafik (Graphics Processing Units, GPU). Masa pemprosesan bagi keseluruhan program telah berkurang hampir 94% daripada masa pemprosesan yang asal dan mengekalkan ketepatan kebolehjadian maksimum dan ukuran dahan. Penyelidikan ini telah mencapai ketepatan bagi pokok filogenetik dan mempunyai ukuran dahan yang bagus dan dapat mengurangkan masa pemprosesan untuk proses pembinaan pokok filogenetik.

GPU BASED FAST PHYLOGENETIC TREE CONSTRUCTION ALGORITHM WITH REDUCE DATASET

ABSTRACT

The tremendous growth of new genomic data, the enhancement and the fusion of genomic data analysis methods and the manipulation of the technological innovations designed for high performance computing had become the main interest of this research. Genomic data analysis; sequence analysis is used to analyse and manipulating the homologous genomic data and phylogenetic tree is one of the method in sequence analysis to construct the evolutionary relationship between the genomic data. However, the construction of a phylogenetic tree required an initial process that is sequence alignment process. This researched had proved that the input genomic dataset must be aligned before the phylogenetic tree construction process took place. Sequence alignment is a process to align the genomic data in finding the similar regions. This is an important process because the raw homologous genomic dataset usually are not standardized and consist of unknown characters. Nowadays, there are large numbers of sequence alignment's programs that available to be employed. Hence, the selection of an ideal program to align the dataset becomes more difficult. Preliminary experiments conducted had proved that best program to align the sequences dataset is MAFFT compared to ClustalW, Kalign, MUSCLE and T-Coffee. The result of sequence alignment is an aligned dataset. The aligned dataset was used as the input dataset for constructing a phylogenetic tree. There are a lot of programs available with various kinds of methods to construct a phylogenetic tree. A comparative study was conducted to compare the methods from a few notable phylogenetic tree construction programs; GARLI, MrBayes, Tree Puzzle and FastTree. Evaluation had shown that FastTree appeared as a program that has many

robust methods to construct a phylogenetic tree such as neighbor-joining method and profile-based method for the arrangement of nodes and taxas position of the tree. Through the experiments to construct a phylogenetic tree, we found that, aligned sequences selection also able to affect the phylogenetic tree construction process and result. Hence, a method was introduced to increase the quality of the aligned dataset; Half-parsimonious. Half-parsimonious method was able to reduce the size of the dataset while keeping the informative sites. This method was able to increase the maximum likelihood score and the branch length of the phylogenetic tree while decreasing the processing time for the construction process. The informative aligned dataset then will be used as the input data for the integration of phylogenetic tree construction's methods. Our experiments shows that the new integration methods able to increase the maximum likelihood scores and the branch length of the phylogenetic tree. However, the processing time of this new integration had increase due to the exhaustive search algorithm implemented in the construction process . Hence, an acceleration method was implemented by using the many-core processors; Graphic Processing Unit (GPU). The processing time for the accelerated phylogenetic tree construction process was reduced almost 94% from the original process while maintaining the accuracy of the maximum likelihood score and the branch length. This research had constructed an accurate phylogenetic tree with a good branch length and lower processing time for the phylogenetic tree construction process.

CHAPTER 1

INTRODUCTION

1.1 Background

Bioinformatics is a scientific interdisciplinary research that been derived from the informatics research area which deals with biological data for data manipulation and information processing (Kasabov, 2014). This bioinformatics field includes the research in storing, retrieving, organizing, managing, analysing and visualizing the biological data, such as signal and image processing data, macromolecular structural data, genomic data and etc.

Sequence analysis process had received vast interest for decades as it has large value in species conservation, genomic structure prediction, disease detection, sequence alignment, phylogenetic tree construction, phylogenetic inference and etc. (Asten et al., 2004). This process is important in finding and analyse the features, function, structure and evolution of the genomic data; nucleotide sequence data, protein sequence data and etc. Figure 1.1 shows the research scopes that involve in sequence analysis process.

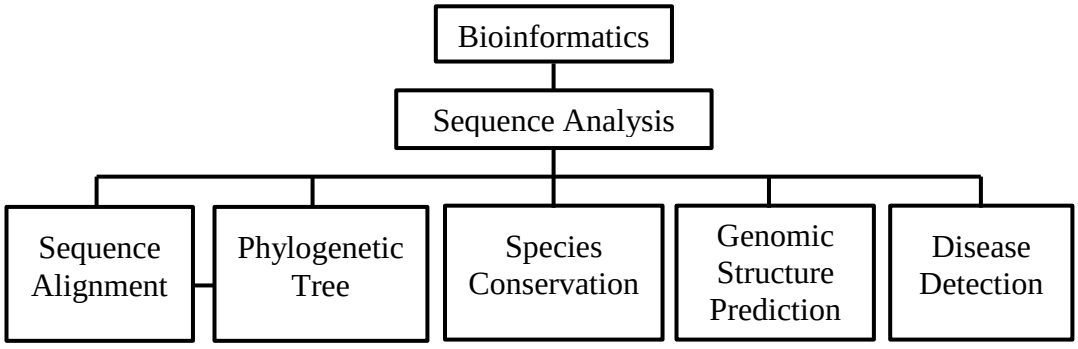

Figure 1.1: Bioinformatics research scope's hierarchy (Asten et al., 2004)

Based on Figure 1.1, there are a correlation between the sequence alignment and phylogenetic tree. The accuracy and performance for constructing a phylogenetic

tree is dependent on the aligned sequences that can be obtained from sequence alignment process (Afiahayati & Hartati, 2010; H. Carroll et al., 2007; Castresana, 2000; Phillips, Janies, & Wheeler, 2000). Hence, this research will cover the sequence analysis areas which include the sequence alignment analysis and phylogenetic tree analysis to construct an evolutionary tree.

Sequence alignment is a process for similarity searching of the genomic dataset by comparing the sequences (Sharma, 2009). This alignment process is necessary for filtering the unknown sequences especially the homologous sequences in order to discover the structure, function and evolution (Blair & Murphy, 2010; Drummond & Rambaut, 2007; Ghosh, Mandal, & Ray, 2015; Isa, Ahmad, Murad, Ismail, & Benkrid, 2014; Mak & Lam, 2003; Pei, Hemani, & Paul, 2011; Suchard & Rambaut, 2009; Yang, 2007). The output of this process will be the complete aligned sequences that are free from deletion, insertion and mutation (Cai, Juedes, & Liakhovitch, 2000; Phillips et al., 2000). The aligned sequences are the best input dataset for determining the evolutionary relation between the sequences involve by constructing a phylogenetic tree (Afiahayati & Hartati, 2010; H. Carroll et al., 2007; Castresana, 2000; Phillips et al., 2000).

Phylogenetic tree construction is a process of constructing an evolutionary relation between the genomic data to explore the history, interrelation and diversity of life on the globe. Charles Darwin had stated that species had spread and transform through evolution at one point. Hence this tree construction of evolutionary tree will be able to classify the connection between the communities, population and species via the point mutation of the genomic data. However, due to the exponential increase of genomic data annually, the computational optimization and enhancement was introduced to analyse the sequences (Alachiotis, Sotiriades, Dollas, & Stamatakis,

2009; Berger, Alachiotis, & Stamatakis, 2012; T. C. Carroll, Ojiaku, & Wong, 2015; Ocana, de Oliveira, Dias, Ogasawara, & Mattoso, 2011).

The computational optimization and enhancement process can be categorized into the algorithm and the technological innovations. This process is important to achieve the scalable, reliable, accurate and high performance of the sequence analysis process (Alachiotis et al., 2009; Berger et al., 2012; Zhou, Liu, Stones, Xie, & Wang, 2011). The algorithm's optimization and enhancement will lead to the simplification of the methods and any mathematical operations involve within the process. There are also the enlightenment of profiling method in handling the big genomic data for data storage and retrieval (Price, Dehal, & Arkin, 2009). The technological exploitation was introduced by the enhancement of the computation process by using high performance computing (Isa et al., 2014). High performance computing is the computational optimization that includes the utilization of many-core processors to overcome the processing time issues while reducing the cost and power consumption during the sequence analysis process and handling the task distribution and scheduling (T. C. Carroll et al., 2015; Isa et al., 2014; Pratas, Trancoso, Stamatakis, & Sousa, 2009; Suchard & Rambaut, 2009). This high performance technology manages to trigger the best performance and increase the computation speed of the high complexity process (Pei et al., 2011; Schmidt, Strimmer, Vingron, & von Haeseler, 2002).

1.2 Motivations and Research Problems

Sequences analysis includes the sequence alignment process and phylogenetic inference process for constructing a phylogenetic tree. Figure 1.2 shows the main framework of sequences analysis to construct a phylogenetic tree.

Figure 1.2: The fundamental framework of phylogenetic tree construction process

Based on Figure 1.2, the phylogenetic tree can be constructed by aligning the homologous genomic data from the database. The genomic data carries great number of gene information and the collection is increasing annually which had affected the sequence analysis process; sequence alignment and phylogenetic tree construction (Flicek & Birney, 2009; D. Yao, Jiang, You, Abulizi, & Hou, 2015; Zierke & Bakos, 2010).

Sequence alignment process is an important process to identify the similarity region of a group of huge genomic data which will determine the function and evolutionary of the sequences (Afiahayati & Hartati, 2010; H. Carroll et al., 2007; Castresana, 2000; Phillips et al., 2000). The aligned sequences will affect the phylogenetic tree construction process by constructing a deep branch tree with the best divergent between species (Li-San et al., 2011). Nowadays, there are a lot of sequence alignment programs with various types of methods that are available and well known. Hence, the main benchmarks was set to determine the ideal sequence alignment process; high position, high percentage of close neighbour, lower percentage of bad split, shortest processing time and maximum likelihood score

(Castresana, 2000). However, not all of the sequence alignment process able to align the big genomic sequences to meet the benchmark score.

Phylogenetic tree construction process includes finding the relation of the genomic data sequences which will classify the ancestor and descendants of the sequences. This construction process withstands a great challenge in data handling due to the abundant of aligned dataset as an input (Blair & Murphy, 2010; Castresana, 2000; Stamatakis, Ludwig, & Meier, 2004; Yang, 2007; Zhang, Wang, Lin, & Feng, 2014). There are many phylogenetic tree construction methods available. However, not all of the methods are suitable for the large dataset handling and able to fulfil the benchmark of constructing a good phylogenetic tree such as; a maximum likelihood tree and less processing time (Blair & Murphy, 2010; Stamatakis et al., 2004). A phylogenetic tree with maximum likelihood score is important to ensure the high divergency of the tree as the classification of species happened at the deepest tree's taxa which result the optimal phylogenetic tree (Kazutaka Katoh, Kuma, & Miyata, 2001; Matsuda, 1996; Stamatakis et al., 2004). The optimal phylogenetic tree, however affect the processing time and decrease the performance of the construction process due to the complex mathematical computation for each method (Stamatakis et al., 2004).

There are the needs of computation process enhancement by exploiting the number of processors and the computer architecture for the complex computation of the phylogenetic tree construction process. The enhancements are for the large genomic data handling and massive number of iterations for each method involve in the maximum likelihood process and tree construction that had caused a limitation on processors' scalability (Alachiotis et al., 2009; Berger et al., 2012; Suchard & Rambaut, 2009; Zahid, Hasan, Khan, & Ullah, 2015; Zhang et al., 2014). Hence, data

distribution, task scheduling and task distributions will able to show an important role in processing the big genomic data, enhancing the iterative computation and have a scalable execution process (Bakos, 2007; Papadonikolakis, Bouganis, & Constantinides, 2009; Zhou et al., 2011).

This research will prove that the improvement of aligned sequences to form the informative dataset will able to construct an accurate and better performance of phylogenetic tree with improvement of the methods the enhancement of the process by using the new technologies innovation.

1.3 Research Questions

i. What are the sequence alignment programs that can accurately aligned sequence without sacrificing the performance and processing time?

ii. What are the phylogenetic tree programs available and methods involve in constructing the high divergency and maximum likelihood of the tree without sacrificing the processing time?

iii. How the data reductions contribute towards the improvement of maximum likelihood and processing time of a phylogenetic tree construction process?

iv. How the implementations of parallel approach in the tree construction process using GPU technology manage to provide a better performance?

1.4 Research Objectives

This research aims to improve the construction of a phylogenetic tree with integrated and enhanced algorithm with reduce dataset. Hence, the research objectives are as follows:

i. To optimize the phylogenetic tree construction process by reducing the aligned sequences for local alignment approach for increasing the tree's accuracy and reducing the processing time.

ii. To construct a maximum likelihood tree using optimizes bifurcation method via profiling approach.

iii. To implement a parallel approach; GPU to enhance the performance of phylogenetic tree construction process.

1.5 Research Scope

This research focuses on the sequence alignment process, local alignment approach on reducing the aligned sequences dataset and constructing an accurate phylogenetic tree using distance-based method, character-based method and profiling method with data load balancing and task scheduling on the GPU technology for faster processing time. Sequence alignment process is important to align the genomic sequences dataset (R. Edgar, 2004; Notredame, 2007). The local alignment approach is used to reduce the aligned sequences dataset for determining the informative sites of the aligned sequences by implementing the dataset reduction methods (Campo et al., 2014).

Distance-based is a method to measure the dissimilarity and optimize the evolutionary distance between the sequences. Whereas, character-based is a method to align the sequences to reach the optimal evolution for the maximum likelihood score of the phylogenetic tree. Profiling method is used for the fast store and retrieves data of the maximum likelihood score and branch length for comparison purposes in constructing the end structure of phylogenetic tree. Finally, GPU technology is used to enhance the processing time of constructing a phylogenetic tree

by exploiting the scalability methods; 1) Load balancing for scalable data distribution and 2) Task scheduling for scalable task distribution, over the threads and processors of GPU.

The limitation of this research involved the genomic dataset which the increasing of junk characters had made some of the MSA programs for sequence alignment process replied an error message and the dataset itself need to be repair manually. The homologous sequences of the dataset also play a big role as an input data for constructing a phylogenetic tree (Isa et al., 2014). The non-homologous sequences will lower the level of correctness for the phylogenetic tree due to the high memory space requirement and the high branch's divergency. This state of using the non-homologous sequences can be called as homoplasy. Homoplasy will cause the false positive result for the tree.

1.6 Research Contributions

i. A method to reduce the amount of data used in constructing a phylogenetic tree. This reduction will reduce the processing time.

ii. A method to optimize the bifurcation process in constructing a phylogenetic tree. This optimization will construct an accurate maximum likelihood tree.

iii. An advance computation for parallel approach in GPU to construct a phylogenetic tree. This enhancement will increase the performance and processing time.

1.7 Thesis Organization

This thesis consists of eight chapters organized as follows:

Chapter 2: Gives overview of the genomic data type, the introduction of the approach and methods sequence alignment and phylogenetic tree, the basic tree topology, the basic model of nucleotide evolution and the high performance computation.

Chapter 3: Discussed the methodology of this research which consist of the theoretical framework and conceptual framework.

Chapter 4: The experimental analysis of the several experiments: 1) Determining a MSA program for aligning the dataset; 2) Determining the phylogenetic tree construction methods.

Chapter 5: The experimental analysis on one of the proposed work: To reduce the size of dataset which by extracting the informative site during the block-segmentation alignment.

Chapter 6: The experimental analysis of constructing a phylogenetic tree using the integration of methods that viewed on Chapter 4.

Chapter 7: The experimental analysis of the experiment on the acceleration of phylogenetic tree construction process using GPU.

Chapter 8: Summarizes and concludes the thesis and also some recommendation on the future works.

CHAPTER 2

PRELIMINARIES AND RELATED WORK

2.1 Introduction

Sequence analysis had received vast interest for decades as it has enormous value in species conservation, structure prediction, disease detection, phylogenetic inference and etc. This research is focusing on the sequence analysis by constructing a phylogenetic tree with enhancements of the construction process to achieve maximum likelihood with high performance computing. Phylogenetic is the sequence analysis study to find the evolutionary of the tree of life that had built the connections between groups of organism. It classifies communities and populations of species of living things. The classification of species is a ubiquitous subject in the genetic sequence analysis process as it becomes more complicated and challenging over the years and the complicacy continue to be increasing dramatically due to new discovery of species genetic specimens.

The phylogenetic tree construction process depends on the dataset, which must be aligned before the analysis. This aligned dataset can be constructed by sequence alignment method and use as the input data for constructing a phylogenetic tree. Sequence alignment is an approach to eliminate the unknown sequences (Morgenstern, Frech, Dress, & Werner, 1998). This sequence alignment will start the dataset searching process by implementing a comparative method to compare the selected sequences with the existing sequence (Sharma, 2009). The outcome of this sequence alignment is the complete sequences that are free from deletion, insertion and mutation.

The enchantments of this phylogenetic tree construction process are included the optimization of the algorithm and high performance computing (Isa et al., 2014). To achieve the ideal sequence alignment for processing the input data for the tree and to construct the maximum likelihood phylogenetic tree with high speed, some preliminaries studies had been carried out.

This chapter covers the background of this research fields: biological data in Section 2.2, basic genomic alignment in Section 2.3, sequence alignment method in Section 2.4, the introduction of phylogenetic tree in Section 2.5,the classification of phylogenetic tree in Section 2.6, the tree topology in Section 2.7, the bifurcation method for evolutionary in Section 2.8, the overview of the available methods in phylogenetic tree construction programs in Section 2.9 and the summary of the phylogenetic tree programs' overview.

2.2 Biological Data: Genomic Data

Every living thing has the same basic type of biological molecules to build their building blocks. Each building block has their own encoding molecules which determine the genetic information and differ from each other. The primary members of biological molecules are Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA) and protein amino acid. Table 2.1 shows the characteristics of these genomic data.

Table 2.1: Characteristics of DNA, RNA and Protein

<table>
<tr><th></th><th>DNA</th><th>RNA</th><th colspan="4">Protein</th></tr>
<tr><td>Shape (s)</td><td>Single shape; double helix and twisted ladder</td><td>Multi shapes</td><td colspan="4">Multi shapes</td></tr>
<tr><td>Structure</td><td>Double-stranded and long chain</td><td>Single-stranded and shorter than DNA nucleotide sequence’s chain</td><td colspan="4">One or more polypeptide chains folded and coiled together</td></tr>
<tr><td>Functions</td><td>Store genetic information and instruction</td><td>Transfer the genetic information for the creation of proteins</td><td colspan="4">Use for support, storage, transportation of other substance, defence against invader and catalytic enzymes</td></tr>
<tr><td>Bases</td><td>(A) - Adenine
(C) - Cytosine
(G) - Guanine
(T) - Thymine</td><td>(A) - Adenine
(C) - Cytosine
(G) - Guanine
(U) - Uracil</td><td>(A) - Ala - Alanine</td><td>(I) - Ile - Isoleucine</td><td>(R) - Arg - Arginine</td></tr>
<tr><td></td><td></td><td></td><td>(C) - Cys - Cysteine</td><td>(K) - Lys - Lysine</td><td>(S) - Ser - Serine</td></tr>
<tr><td></td><td></td><td></td><td>(D) - Asp - Aspartic Acid</td><td>(L) - Leu - Leucine</td><td>(T) - Thr - Threonine</td></tr>
<tr><td></td><td></td><td></td><td>(E) - Glu - Glutamic Acid</td><td>(M) - Met - Methionine</td><td>(V) - Val - Valine</td></tr>
<tr><td></td><td></td><td></td><td>(F) - Phe - Phenylalanine</td><td>(N) - Asn - Asparagine</td><td>(W) - Trp - Tryptophan</td></tr>
<tr><td></td><td></td><td></td><td>(G) - Gly - Glycine</td><td>(P) - Pro - Proline</td><td>(Y) - Tyr - Tyrosine</td></tr>
<tr><td></td><td></td><td></td><td>(H) - His - Histidine</td><td>(Q) - Gln - Glutamine</td><td></td></tr>
</table>

Nowadays, biology database had increased tremendously. Due to the increases of findings on the new species, GenBank's databases had faced rapid growth. In the last few years, there are a lot of genomic data had been extracted from new findings of biological specimens (Zierke & Bakos, 2010). Figure 2.1 shows the current graph of growth of GenBank as the proof that the database had become bigger and complex.

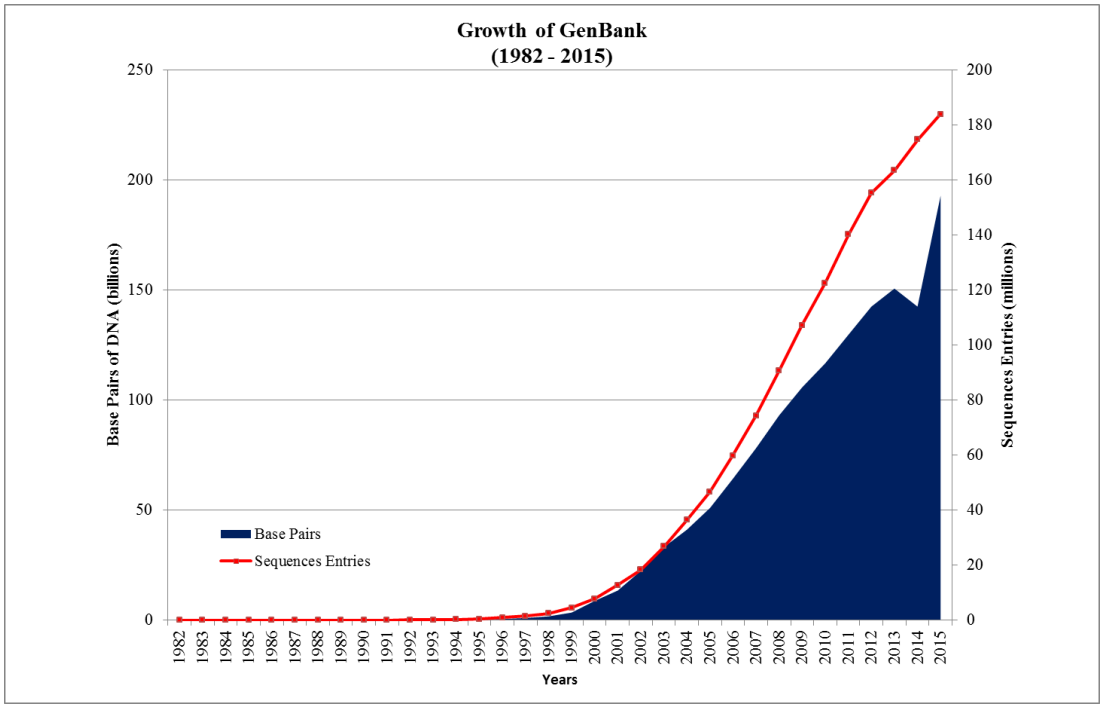

Figure 2.1: Growth of GenBank from 1982 – 2015 (GenBank, 1982)

Based on the Figure 2.1, the graph had shown the rapid growth of the sequence number in the GenBank, yearly. The growth process of the graph had implicated the sequence alignment process since to construct the good alignment, the input dataset must be a set of sequences which have higher similarity, called as homologous sequences (Isa et al., 2014). Homologous sequences is the set of sequences which have high similarities due to the shared of ancestry whereas, the non-homologous sequences shown the characteristic of having distant similarity

(Campbell, 2005). The non-homologous sequences are also been able to be aligned to find the similarities within the sequences. However, the result of the non-homologous sequences will return the false positive aligned sequences (Pirovano & Heringa, 2008).

The false positive alignment is the aligned sequences which produced with a lot of junk dataset and have no similarity at all, but roughly will show the similarity between the sequences (Pirovano & Heringa, 2008). The homologous dataset can be obtained from any trusted available digital databases that consist of gold standard database benchmarks. Some of the well-known benchmarks are: Benchmark Alignment Database (BAliBASE), Protein Reference Alignment Benchmark (Prefab), Homologous Structure Alignment Database (HOMSTRAD), National Institutes of Health Geanetic Sequence Database (GenBank) which also consist of DNA DataBank of Japan (DDBJ) and Europian Molecular Biology Laboratory (EMBL) and etc. Table 2.2 shows the available benchmarks that consist of homologous sequences provided with the type of sequences present and links to reach the resources' files.

Table 2.2: The availability of benchmarks and type(s) of sequences accessible

	Type(s) of Sequences	**Resource**
BAliBASE	Protein	http://www-bio3d-igbmc.u-strasbg.fr/balibase/
HOMSTRAD	Protein	http://tardis.nibio.go.jp/homstrad/
GenBank	DNA, Protein	http://www.ncbi.nlm.nih.gov/genbank/
EMBL	DNA, Protein	http://www.ebi.ac.uk/
DDBJ	DNA	http://www.ddbj.nig.ac.jp/

2.3 Genomic Dataset Alignment

Aligned dataset sequences are the important input in constructing a phylogenetic tree. However, the raw genetic molecular dataset were widely known as the unaligned sequences which are not standardize and sometimes consist of unknown character. Based on research by Asten et al. (2004), the sequences analysis of phylogenetic tree can be determine by using an unaligned dataset and aligned dataset. However, not many programs can use raw dataset to construct a phylogenetic tree.

In early introduction of genomic dataset alignment, two types of approaches were introduced, that are global alignment and local alignment (Lassmann & Sonnhammer, 2005; Needleman & Wunsch, 1970; Smith & Waterman, 1981). The global alignment is focusing on the optimization of the dataset whereas the local alignment is focusing on the segmentation of the sequences which consist of the important genetic information known as the informative's segments (Morgenstern, 1999).

2.3.1 Global Alignment: Approach for Optimization

Needleman-Wunsch algorithm for sequence alignment had highlighted the usage of Dynamic Programming (DP) with global alignment to discover the alignment score to identify the match, mismatch (point of mutations) and gaps (insertion and deletion) within the PA (Needleman & Wunsch, 1970). Global alignment is an approach that enable the optimization of similarity searching which maximizing the PA similarity which best describe the relations between sequences (Mohsen, Zainol, Salam, & Husain, 2007; Needleman & Wunsch, 1970). However, global alignment also have its own limitation as there will be large gaps insertion within the sequences which will lead to the waste of time and power consumption (Pirovano & Heringa, 2008; Siddesh, Srinivasa, Mishra, Anurag, & Uppal, 2015; D. Yao et al., 2015). This

type of sequence alignment has higher chances of getting higher accuracy but, lack of efficiency and low significant similarity.

2.3.2 Local Alignment: Approach for Informative Sequences

Local alignment is the type of sequence alignment which only align a conserve region which means the sequences that are closely related and eliminate the unnecessary part in particular sequences (Mohsen et al., 2007; Morgenstern et al., 1998; Smith & Waterman, 1981). Local alignment can be found in Smith-Waterman algorithm (Smith & Waterman, 1981). Local alignment approach is the suitable approach to start the sequence analysis as the approach will maximized the alignment score locally and only keep the informative segments that will contribute in the speed up of the sequence analysis process (Besharati & Mehrdadjalali, 2014; Chairungsee, 2014; Morgenstern, 1999; Pirovano & Heringa, 2008).

Local alignment approach works for homology sequences, which is an appropriate approach to begin the sequence alignment process that composed of high significant similarity. The type of alignment suitability for convergence is also depending on the range of sequence homogeneity (Pirovano & Heringa, 2008). Figure 2.2 shows the alignments' details which are important for entire sequence analysis purposes.

Figure 2.2: Alignment types approach

2.4 Sequence Alignment Methods

Sequence alignment methods can be classified into Pairwise Alignment (PA) and Multiple Sequence Alignment (MSA). PA is a method which can align two sequences at a time, whereas MSA is a method that enables the alignment of three or more sequences at a time. Both of these alignment methods will produce a set of information-rich aligned sequences dataset (Siddesh et al., 2015). Figure 2.3 shows the basic order of sequence alignment process. These steps become the standard benchmark for the entire sequence alignment program.

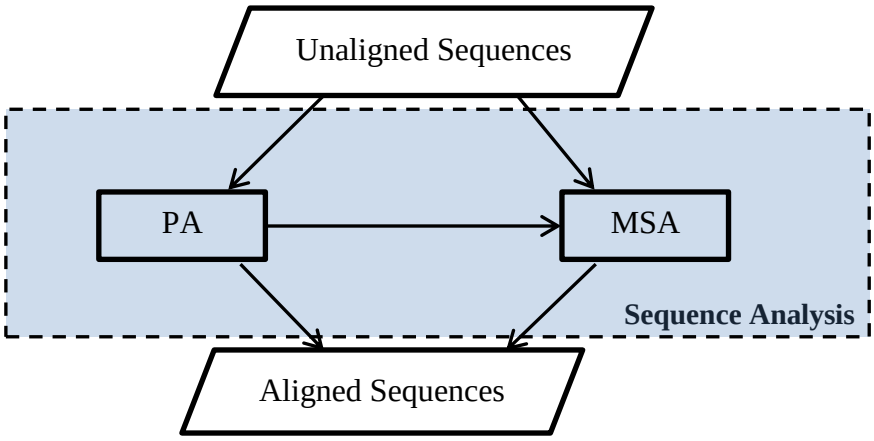

Figure 2.3: Basic order of sequence alignment process (Notredame, 2007)

2.4.1 Pairwise Alignment (PA) Approach

PA is a basic method on aligning two sequences introduced by Needleman and Wunsch (1970) and the research being pursue by Smith and Waterman (1981). This alignment method is well-known at first with the implementation on protein substances and later on being implemented on DNA and RNA. The early PA alignment method was focused on the calculation process, which emphasize the usage of Dynamic Programming (DP). But, later on, there are some enhancements were invented, that are matrix transition probability and grading function (Sharma, 2009). There other enhancement was the scoring function which the sequences being

computed and do the similarity searching with appropriate scoring method benchmarks such as BLOSUM and PAM for protein amino acid sequences. This scoring function can substitute the grading function (Mohsen et al., 2007; Sharma, 2009). This alignment method is the best method which offers an accurate aligned result, though it is not sufficient for more than two sequences as it is time consuming (Siddesh et al., 2015).

Current technologies had discovered a lot of methods to align the unaligned sequences besides using PA. The enhancement of PA is Multiple Sequence Alignment (MSA) which able to discover more similarities of a large group dataset in details compare to PA (Lassmann & Sonnhammer, 2005). However, PA also becomes the building block of MSA as the alignment method becomes the sub-steps in MSA method (refer Figure 2.3).

2.4.2 Multiple Sequence Alignment (MSA)

MSA is a sequence searching method which able to align three and more sequences. Since mid-1980s, MSA had become the selected method to align the selected sequences, as it can reduce the time consumption as a lot of sequences will be aligned at the same time. However, MSA offer approximate and not optimal alignment result.

Alignment of sequences can be analysed by going through some steps that also include the notable method that is a matrix-based method. However, due to the increasing of large dataset each year, simultaneously with the additional of processing time, there also other methods were introduced. Heuristics and probabilistic are the best methods to be practice due to the large dataset handling. Nevertheless, nowadays, this method also had faced some issues and obstacles which

had brought it both down to the matrix-based level. Researcher needs to start to discover this interdisciplinary limitation by turn their research towards the hardware architecture of their processing machine.

According to (Notredame (2007); Pirovano and Heringa (2008)), computational MSA is not a simple challenge. Even, MSA also have early method which still using PA as the initial step to get the pairwise evolutionary distance, especially while using distance method. The score of the sum of pairs of MSA also the extend of DP (Konagurthu & Stuckey, 2006).

To determine the multiple sequence alignment, there are two main methods that are scoring method and clustering method. The scoring method is implementing the global alignment by using DP which the MSA is determined by using Sum-of-Pairs (SP) grade (Sharma, 2009). The scoring scheme is implemented by the PA method which had integrated with MSA method. The scoring scheme is divided into two, that are; 1) matrix-based, 2) consistency-based (Notredame, 2007). The example of MSA programs which are using matrix-based were ClustalW (Larkin et al., 2007; Thompson, Gibson, Plewniak, Jeanmougin, & Higgins, 1997), MUSCLE (R. Edgar, 2004; R. C. Edgar, 2004) and Kalign (Lassmann & Sonnhammer, 2005). The example of consistency-based MSA was T-Coffee (Edgar & Batzoglou, 2006; Notredame, 2007; Wallace, Blackshields, & Higgins, 2005). According to Notredame (2007), the consistency-based is more accurate but have higher CPU times compare to other methods.

The clustering method can be divided into two groups that are progressive method and iterative Method. Iterative method is the result of enhancement in progressive method. Regarding to (Notredame, 2007), while progressive alignment is

the greedy heuristic assembly algorithm which the algorithm had resulted the estimation of unaligned sequence using guide tree and running the MSA via PA method (refer Figure 2.3).

2.4.3 Overview Programs of Multiple Sequence Alignment (MSA)

There are several well-known programs for MSA process to construct the aligned sequences; ClustalW, Kalign, MAFFT, MUSCLE and T-Coffee. Almost all of the programs are flexible, available and notorious. The principal of these programs is to aim for an accurate, an efficient and rapid search of the raw dataset in order to construct the accurate or approximate aligned sequences. The result of MSA can be use as the input data to construct a phylogenetic tree.

2.4.3 (a) ClustalW

Clustal is a program that used the weight matrix calculation method (Saitou & Nei, 1987; Thompson, Higgins, & Gibson, 1994). Clustal is a matrix-based program, that can align a medium size of data protein, DNA and RNA, which calculate the best score with limitation of 500 sequences or 1MB of data (EMBL-EBI, 2013; Larkin et al., 2007; Liu, Linder, & Warnow, 2011; Notredame, 2007). Clustal will result the findings of similarities, identities and differences of selected sequences.

Clustal program was enhanced to become more sensitive in handling the alignment process of high divergence sequences such as the improvement on the sequences weighting, gap penalties, position specification and also the addition of other weighted method, Neighbor Joining (NJ) method (Higgins & Sharp, 1988; Larkin et al., 2007; Thompson et al., 1994). Clustal program also known as the optimizer of other MSA programs, appear after early 1990s (Kazutaka Katoh, Asimenos, & Toh, 2009; Wallace et al., 2005).

Clustal had been introduced as ClustalW in 1994, and there are enhancements, until the recent version, ClustalW2. The enhancement had been made because of the high computational cost even though the alignments were accurate. The accuracy came from the process of align and realign of each sequences which will lead to the usage of a lot of memory and computation time. ClustalW is focusing on scoring scheme and weighting scheme which had combined the usage of Neighbor-Joining (NJ) and UPGMA (Larkin et al., 2007). Figure 2.4 shows the framework of ClustalW.

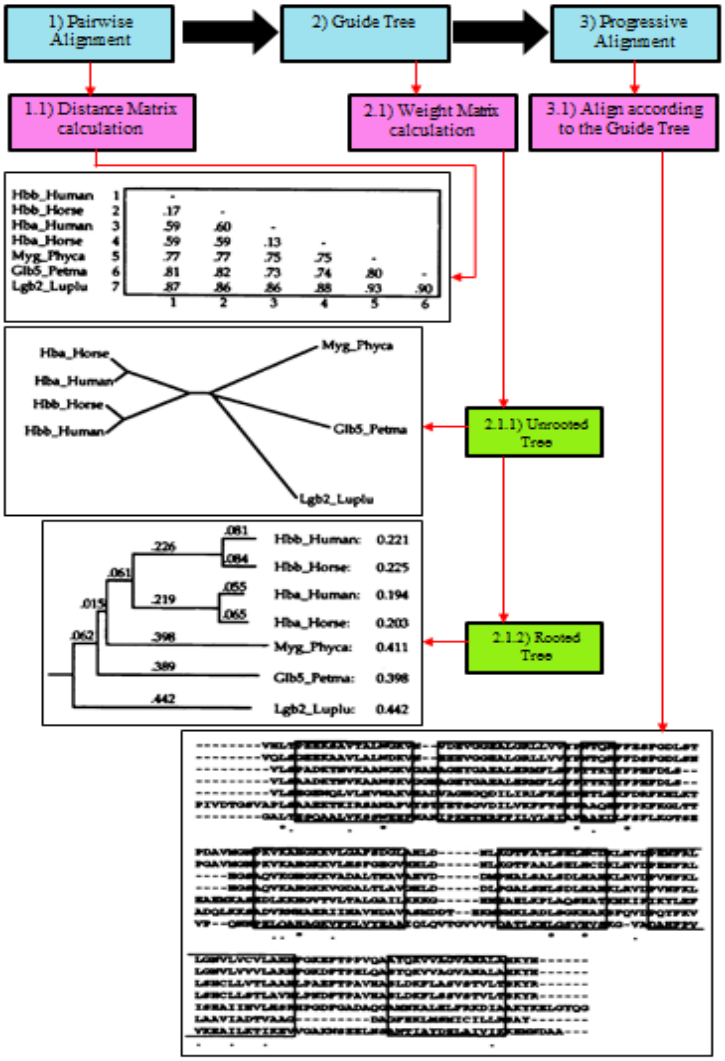

Figure 2.4: ClustalW mainframework (Thompson et al., 1994)

Based on Figure 2.4, there are 3 main stages for aligning the sequences in ClustalW. Based on the Figure 2.4, the stage (1) is PA method to determine the (1.1) distance matrix. The divergence of the selected sequences then were counted and the gaps penalties were introduced (gaps opening and gaps extension). The higher the divergence sequence, the large gaps extension will be.

Stage (2) is a guide tree construction. This is the improvement that made by Thompson et al. (1994) with the implementation of NJ to calculate the (2.1) weight matrix and resulted in (2.1.1) unrooted tree (Larkin et al., 2007; Thompson et al., 1994). The advantages of NJ during the usage of less similarity sequences compare to the former UPGMA was being able to construct a better length estimation (Thompson et al., 1994). The root was determined by determining an out-group of the species involved and the (2.1.2) rooted tree will be generated. In guide tree, the enhancement was on the calculation of the branch length. The slow calculation of branch length will generate an accurate aligned sequences, whereas, the fast calculation will generate an approximate aligned sequences (EMBL-EBI, 2013). The last stage is (3) progressive alignment method where the sequences had been aligned according to the guide tree's branching order progressively.

There are some enhancements to date on ClustalW as ClustalX and ClustalW-MPI had been introduced. ClustalX was developed to emphasize the program's result representation, visually (Thompson et al., 1997). ClustalX 2.0 which being introduced and contribute towards improvement of UPGMA (Larkin et al., 2007). ClustalW-MPI is one of the program that been developed to enhance the ClustalW program with the catalyst of high performance processors. This MPI implemented program was run in distributed and parallel enable processors without changing the algorithm and the execution time was able to be reduce (Li, 2003). The optimization

of ClustalW also been introduced on shared-memory implementation by using Silicon Graphics International (SGI) and Posix threads (Pthreads) (Dmitri Mikhailov, Haruna Cofer, & Gomperts, 2001; Li, 2003).

2.4.3 (b) T-Coffee

Tree-based Consistency Objective Function For alignment Evaluation (T-Coffee) is the progressive MSA method which become one of the convincing programs in MSA implementation (Notredame, 2007; Wallace et al., 2005). Throughout the year after being invented, T-Coffee has faced a lot of library's enhancement by Notredame's lab starting from year 2000. The enhancements are due to the some common limitation on sequence analysis process such as the increasing of dataset each year. T-Coffee is the consistency-based method of progressive alignment that can align and combine both of PA method, the local and global alignment (Notredame, 2007). This combination had returned some highly positive good result. However, the computational process of T-Coffee took a long time. This rising issue will lead to the time consuming problem (Lassmann & Sonnhammer, 2005).

The early enhancement of T-Coffee were 3D-Coffee in 2004 which had increased the accuracy, fast and simple (Edgar & Batzoglou, 2006; Wallace et al., 2005). Then, other enhancement had taken place such as M-Coffee in 2006 to estimate the consensus alignment, implementation on graph based in 2008, R-Coffee in 2008 for RNA, Cloud-Coffee in 2010 for parallel, cloud and non-GPU, Web-Services and 3D structures in 2011 and the latest enhancement in 2012 was GPU MSA. Figure 2.5 shows the basic framework of T-Coffee.

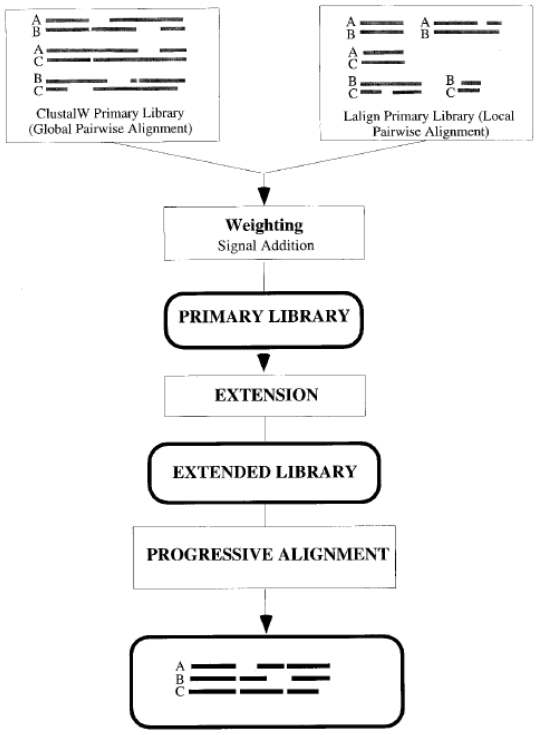

Figure 2.5: The framework of T-Coffee (Notredame, Higgins, & Heringa, 2000)

2.4.3 (c) MAFFT

MAFFT is one of the MSA programs that are based on progressive alignment method. MAFFT was enhanced by optimizing the Weighted Sum of Pairs (WSP) and the similarity comparison is identified by implementing Fast Fourier Transform (FFT). The enhancement was made in order to achieve low CPU time and high accuracy of aligned sequences in large size of dataset and less homogenous sequences (K. Katoh, Misawa, Kuma, & Miyata, 2002; Wallace et al., 2005).

MAFFT is suitable to be used with small dataset and this program can be run with progressive method or iterative refinement method (K. Katoh et al., 2002; Liu et al., 2011). The progressive method is the fast method known as FFT-NS-2 and the iterative refinement method is the accurate method known as FFT-NS-i (EMBL-EBI, 2013; K. Katoh et al., 2002). The methods implement in MAFFT is comparable to ClustalW and T-Coffee (K. Katoh et al., 2002). Figure 2.6 shows the comparison